# Node Pruning for Improved Neural Network Design

Ashiquzzman Akm*, Dong Su Lee*, Sang Woo Kim*, Lin Van Ma*, Um-Tae Won** and Jin Sul Kim*

**Abstract**

Deep learning neural networks are getting more attentions each day for their accuracy and improved structural prediction. However, for improving the accuracy, the whole network size is getting bigger as well, which in terms of computation, often gives disadvantages to the users for the necessity of high configuration computers. Although many researches had been focused on the accuracy improvement, very little research has been focused on the network size reduction and filter pruning. In this research, several network node pruning optimization have been attempted to improve the whole network size with respect to steady performance.

**Key words**

*Neural Network, Deep Learning, Node Pruning, Optimization,*

## Ⅰ. Introduction

Network Pruning or Neural Network node removal without reducing accuracy is not an recent idea. The whole Idea was first proposed by the LeCun at 1980. The main idea was described by the “Optimal Brain Damage” [1]. The main idea is for this method is to rank the node based on their performance in each time and cut off the least useful node in each layer based on lowest scores achieved in one single evaluation. However, the idea was not popular in the time because of the lack of mathematical metrics to rank the nodes based on their values. Now the modern Neural Networks are getting bigger with new versions. Although the accuracy is improving to a remarkable rate, the bigger network is making the neural network an expensive computational resources in both training and testing situations. The VGGnet proposed and trained by the Oxford’s Visual geometry group has over 138,357,544 trainable parameters [2]. This model makes the image classification in large scale very efficient with error late less than 7%. However, training this model containing large trainable parameters takes very big computation resources such as memory and processing. So, Deploying this models in a remote machine with limited resources is not feasibly efficient. The whole model of remote resources is small data and small power consumption. So, the idea for filter or node pruning are getting more attention nowadays again for such applications such as Internet of Things (IoT) devices and Mobile Edge Computing (MEC).

In this paper, the idea of neural network pruning for optimized deep-learning neural network (DNN) to classify hand written digit was explored. the proper process for node based on layer ordering was studied. the proposed method for compacting DNN showed promising compression rate and steady accuracy for given datasets.

* School of Electronics and Computer Engineering, Chonnam National Uniersity, Gwangju, South Korea
** Department of Information and Communication Engineering, Chosun University, Gwangju, South Korea

## II. Background Study

In the paper [3], ranking based method was proposed to explore sparsity in activations for network pruning. Rectified Linear Unit (ReLu) activation function imposes sparsity during inference, and average percentage of positive activations at the output can determine importance of the neuron. Average Percentage of Zeros (APoZ) is useful to estimate saliency of feature maps in given layer. This Idea can be used to make the order of ranking of a filters in each layers. The main idea of filter or node pruning can be visualized in the Fig. 1. The middle node has been removed and slashed according to the ranking of the filter APoZ algorithms.

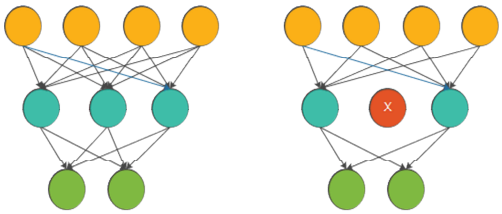


Fig. 1. Node Pruning in Neural Network

## 1 Proposed Method And Result

The main proposed method for the study was divided into several steps. In the initial stage the neural network architecture was defined. The Deep fully connected architecture was decided based on the trail and error on a subset of main dataset. As the dataset for the proposed study has limited target label, the initial size of the neural network was not make exceptionally big. The whole neural network was designed to have 5 (Five) hidden layers with several ranges of initial neurons in the side. the main input layer has the initial neuron to 550 and the later layer were equiped with initial 375 nodes in each layer. then the output layer were fixed based on the input image dimensions.

All the proposed models were constructed in the Python based open-source library Keras and Pytorch, a python based torch library for deep-learning modeling. All the models described in proposed section were trained in the computer with Inter i7 8 core CUP with 8 Gb DDR4 RAM. The Nvidia CUDA library was used to speed up the traning process with the help of single instance to Nvidia Geforce 1050-ti GPU. The above mentioned GPU has 8 GB of VRAM. All the learned weights are being saved in the Hadoop file systems (HDFS) for safe keeping.

The result for the filter pruning was very beneficial for small scale deployment of neural networks. The first all network were separately trained till the accuracy improving stopped at all. This process was monitored with automation based on accuracy improvement monitoring. All of the models then was pruned based on APoZ ranking of the layer. The accuracy of all the networks were remained same but the filter removal was different based on dataset.

The open MNIST handwritten digit dataset and the breast cancer dataset were selected for the studied. Both of the dataset are very different in nature. the MNIST [4] digit dataset have 28x28 image information of handwritten digit data and the Breast cancer dataset has around 600 image features or continuous values in a numerical data.

The Neural Network based in MNIST digit had initial 13 million connection and most of the digits datasets were then output based on the one hot encoding output. the whole test rate error of the net was 2.29% and the removed neural net after pruning had 6 milloin connection with same error rate. The model with breast cancer initially acted will with 1.2% test rate with 11 million connections and later the digit and the later pruned model had 5 million connections.

## IV. Conclusion

This paper proposes a novel solution for IoT based deep learning neural network optimization. The IoT based application solution sometimes needs to be both memory

and computation resource efficient. Deep learning models are often gets very big in model size with high and precise accurate classification. This creates difficulty for implementing the application specially developed for remote resource based solutions. The whole mode pruning was done based on the APoZ ranking of the node after traning and the whole network performance was remained same after pruning the network to half of their original size without compromising the accuracy.

## Acknowledgement

This work was supported by Institute for Information &communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2018-0-00691, Development of Autonomous Collaborative Swarm Intelligence Technologies for Disposable IoT Devices). Besides, this research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-2016-0-00314) supervised by the IITP(Institute for Information & communications Technology Promotion). Furthermore, this research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (MEST)(Grant No. NRF-2017R1D1A1B03034429)

## Reference

[1] Y. LeCun, J. S. Denker, and S. A. Solla, “Optimal brain damage,” in Advances in neural information processing systems, 1990, pp. 598–605.

[2] A. Rosebrock, “Imagenet: Vggnet, resnet, inception, and xception with keras,” Mars, 2017.

[3] H. Hu, R. Peng, Y. Tai, C. Tang, and N. Trimming, “A data-driven neuron pruning approach towards efficient deep architectures. arxiv preprint,” arXiv preprint arXiv:1607.03250, 2016.

[4] LeCun, Yann, Corinna Cortes, and C. J. Burges. "MNIST handwritten digit database." AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist 2 (2010).

[5] Wolberg, W.H., & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193 –9196.